Natural Language Processing and XML Retrieval

Alan Woodley

School of Software Engineering and Data Communications

Queensland University of Technology

Supervisor: A/Prof Shlomo Geva

XML, Information Retrieval, Natural Language Queries

Background 1

- IR systems return relevant documents to users
- XML-IR systems return relevant *portions* of documents to users
- XML documents separate content and structure
- To user XML-IR systems users must specify content and structure requirements

Background 2

- Usually XML-IR queries written in formal language such as NEXI:

  //article[about(.//sec, XML-IR)]//p[about(.,NLP)]

- But formal languages are:
  - Too difficult to use (even for XML-IR experts)
  - Too closely bound to physical constraints of collection (need to know article, sec, p etc.)

Aims, Research Question

In comparison natural language queries are

- Intuitive (People use natural language everyday)
- Not bound to document's physical structure (structure expressed at conceptual level)

  "I want paragraphs describing natural language queries in articles which sections about Information retrieval"

- Solution: Add a natural language interface to an existing XMl-IR system

Method

1. **ID user information needs –** Detect structural and content needs
2. **Analysis structure –** Convert conceptual names (section) to tags (sec)
3. **Analysis content –** Extract key terms and phrases
4. **Format NEXI query –** Combine structure and content

Results 1

- Present retrieval performance of 3 approaches (*Hassler*, *Tannier*, *Woodley*) from INEX's 2005 NLP Track
- Each approach converted NLQs to formal language (NEXI) queries
- Converted NEXI input into backend XML-IR system (GPX)
- Approaches also compared baseline (manually constructed NEXI expressions)
- NLP approaches comparable to - or outperform - baseline

Results 2

	Baseline	*Hassler*	*Tannier*	*Woodley*
nxCG[25]				
Strict	**0.1578**	0.1378	0.1378	0.1378
Gen	**0.2885**	0.3814	0.2693	0.2859
ep-gr				
Strict	**0.077**	0.074	0.0775	0.0755
Gen	**0.1324**	0.1531	0.1064	0.1051

Table 1: SSCAS Retrieval Performance

	Baseline	*Hassler*	*Tannier*	*Woodley*
nxCG[25]				
Strict	**0.0662**	0.0662	0.0662	0.0913
Gen	**0.1081**	0.1046	0.1004	0.11
ep-gr				
Strict	**0.0274**	0.0267	0.0304	0.0267
Gen	**0.0272**	0.0287	0.0298	0.0311

Table 2: SVCAS Retrieval Performance

	Baseline	*Hassler*	*Tannier*	*Woodley*
nxCG[25]				
Strict	**0.1267**	0.1133	0.1133	0.1133
Gen	**0.2531**	0.2815	0.3051	0.2446
ep-gr				
Strict	**0.0383**	0.0338	0.0363	0.034
Gen	**0.0608**	0.0641	0.0682	0.0632

Table 3: VSCAS Retrieval Performance

	Baseline	*Hassler*	*Tannier*	*Woodley*
nxCG[25]				
Strict	**0.1267**	0.1267	0.1867	0.1644
Gen	**0.2281**	0.2456	0.2572	0.2136
ep-gr				
Strict	**0.0454**	0.0372	0.0418	0.0483
Gen	**0.0694**	0.074	0.0799	0.0742

Table 4: VVCAS Retrieval Performance

Conclusions

- NLQs perform comparably to formal language and therefore are a viable alternative
- NLP and IR communities should investigate other mutually beneficial opportunities